PATENT ABSTRACTS OF JAPAN

(11)Publication number : **08-161348**
(43)Date of publication of application : **21.06.1996**

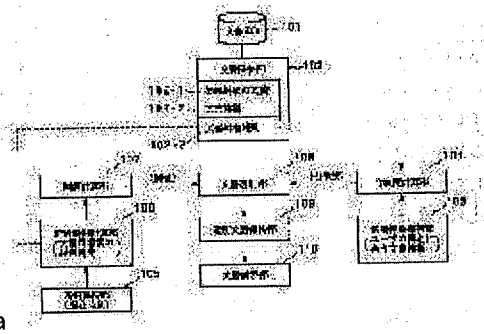
(51)Int.Cl. **G06F 17/30**
G06F 17/21

(21)Application number : **06-298237**
(22)Date of filing : **01.12.1994**

(71)Applicant : **CANON INC**
(72)Inventor : **UEDA TAKANARI**
HIROTA MAKOTO
ITO SHIRO
SHIBATA SHOGO
IKEDA YUJI
FUJITA MINORU

(54) DOCUMENT FILTERING METHOD AND DOCUMENT PROCESSOR

(57)Abstract:
PURPOSE: To provide a document filtering method and a document processor which filters a document by considering not only coincidence of the contents of the document and user's interest but the newness of the document.
CONSTITUTION: In the document filtering method and the document processor which filters and presents a received document, the document which should be presented is sorted by a previously set identifier and the passing time of the received document (108) and the sorted document is presented (110). At the time of sorting the representing document, the identifier of the received document and the previously set identifier are calculated (104) and based on the comparison between a threshold set corresponding to the passing time and the coincidence, whether to present the received document or not is judged. In another way, a score is given to the received document corresponding to the coincidence and the passing time and based on the comparison between the score and the prescribed threshold value, whether to present the document or not is judged. At this time, the passing time is from the preparation of the document or the reception of it, or both from preparation and from reception.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平8－161348

(43)公開日　平成８年(1996)６月21日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30					
17/21					
		9194－5Ｌ	Ｇ０６Ｆ　15/ 403	３５０　Ａ	
		9288－5Ｌ	15/ 20	５７０　Ｎ	
		9288－5Ｌ	Ｇ０６Ｆ　15/ 20	５８６　Ｂ	

審査請求　未請求　請求項の数13　ＯＬ　（全７頁）　最終頁に続く

(21)出願番号　特願平6－298237

(22)出願日　平成６年(1994)12月１日

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号

(72)発明者　上田　隆也
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　廣田　誠
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　伊藤　史朗
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(74)代理人　弁理士　大塚　康徳　（外１名）

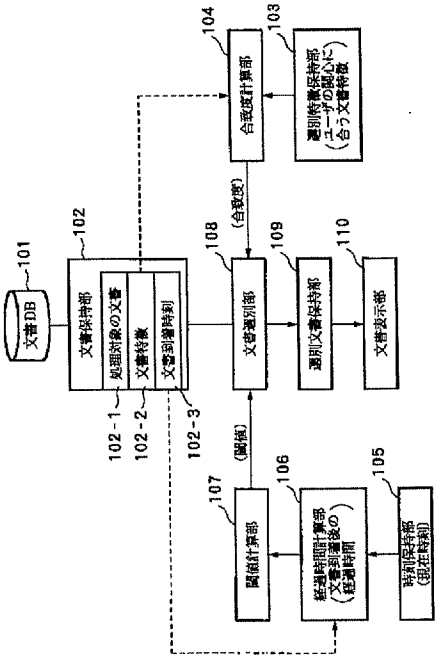
最終頁に続く

(54)【発明の名称】　文書フィルタリング方法及び文書処理装置

(57)【要約】

【目的】　文書の内容とユーザの関心との合致度だけでなく、文書の鮮度（新しさ）も考慮してフィルタリングを行うようにした文書フィルタリング方法及び文書処理装置を提供することを目的とする。

【構成】　受信した文書をフィルタリングして提示する文書フィルタリング方法及び文書処理装置において、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別（１０８）し、該選別された文書を提示（１１０）する。前記提示文書の選別では、受信文書の識別子と前記予め設定された識別子との合致度を計算（１０４）し、前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する。または、前記合致度と経過時間とに応じて受信文書にスコアを与え、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する。ここで、前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方である。



1

【特許請求の範囲】

【請求項１】 受信した文書をフィルタリングして提示する文書フィルタリング方法において、
予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別し、
該選別された文書を提示することを特徴とする文書フィルタリング方法。

【請求項２】 前記提示文書の選別は、
受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、
前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する工程とを含むことを特徴とする請求項１記載の文書フィルタリング方法。

【請求項３】 前記閾値は更に文書数に応じて設定されることを特徴とする請求項２記載の文書フィルタリング方法。

【請求項４】 前記提示文書の選別は、
受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、
前記合致度と経過時間とに応じて受信文書にスコアを与える工程と、
前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する工程とを含むことを特徴とする請求項１記載の文書フィルタリング方法。

【請求項５】 前記識別子は特徴ベクトルで表わされ、
前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされることを特徴とする請求項２または４記載の文書フィルタリング方法。

【請求項６】 前記特徴ベクトルは、複数のキーワードまたは文から成ることを特徴とする請求項５記載の文書フィルタリング方法。

【請求項７】 前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方であることを特徴とする請求項１または２または４記載の文書フィルタリング方法。

【請求項８】 受信した文書をフィルタリングして提示する文書処理装置において、
受信文書を作成からの経過時間及び／又は受信からの経過時間と共に保持する文書保持手段と、
予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別する選別手段と、
該選別された文書を提示する提示手段とを備えることを特徴とする文書処理装置。

【請求項９】 前記選別手段は、
受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、
前記経過時間に対応する閾値を設定する閾値設定手段と、
前記閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含むことを特徴とする請求項８記載の文書処理装置。

2

【請求項１０】 前記閾値設定手段は、更に文書数にも応じて閾値を設定することを特徴とする請求項９記載の文書処理装置。

【請求項１１】 前記選別手段は、
受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、
前記合致度と経過時間とに応じて受信文書にスコアを与えるスコア計算手段と、
前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含むことを特徴とする請求項８記載の文書処理装置。

【請求項１２】 前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされることを特徴とする請求項９または１１記載の文書処理装置。

【請求項１３】 前記特徴ベクトルは、複数のキーワードまたは文から成ることを特徴とする請求項１２記載の文書処理装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は文書処理装置に関し、特にユーザのもとに入ってくる文書のうちユーザが関心を持つ文書を選別し、その結果を出力する文書フィルタリング方法及び文書処理装置に関するものである。

【０００２】

【従来の技術】昨今、記録媒体の大容量化と低価格化、また、ワードプロセッサの普及などによって、電子化された文書の量が増大している。さらに、ネットワークの整備が進み、電子メールや電子ニュースなどのように、メディアを介して直接ユーザのもとに届く電子化文書の量も増えている。このため、ユーザが処理できる量を越えた文書が入ってくるようになり、本当に必要な情報が不要な情報の中にも埋もれてしまうという、いわゆる「情報洪水」が問題になってきている。

【０００３】この問題への対応策として、ユーザが関心を持つであろう文書のキーワードを自動的に選別する「文書フィルタリング」の技術が用いられるようになってきた。「文書フィルタリング」では、予め文書に対するユーザの関心をキーワードや文書の内容として設定しておき、この設定と送られてきた文書の内容を比較して、一致がある場合にはその文書をユーザに見せ、一致していないときにはその文書をユーザに見せないという制御をしている。すなわち、「文書フィルタリング」の技術によって、ユーザは自分にとって関心のないような文書を最初から見なくてすむようになり、時間を有効に利用できるようになった。

【０００４】

【発明が解決しようとする課題】しかしながら、フィル

タリングされた文書を定期的にユーザが見ている間は問題ないが、ユーザがしばらく文書を見ることができずにフイルタリングされた文書が溜まってしまったような場合、文書数が多くなり、結局重要な情報を見落としてしまう可能性が出てくる。

【０００５】従って、重要な情報を見落とさないための対策として、フィルタリングされた文書をユーザの関心のあるキーワードや文書の内容との合致の度合（合致度）の順に整列して表示し、合致度の高いものから見ることができるようにすることも行われるが、この場合は、最新の必要文書が後に来る等、文書の「鮮度」が考慮されていないことに問題がある。

【０００６】電子メールのように日々ユーザのもとに届くような文書については、鮮度が重要で、一般に「新しい文書ほど情報の価値がある」ということが言える。届いたときには重要な情報であっても、時間の経過と共に重要度が下がる文書も多い。しかし、文書フィルタリングにおいて単に文書の内容だけを考慮したのでは、こうしたことが反映されない。

【０００７】従来装置には、以上のような種々の問題があった。本発明は、上述した従来の課題を解決し、文書の内容とユーザの関心との合致度だけでなく、文書の鮮度（新しさ）も考慮してフィルタリングを行うようにした文書フィルタリング方法及び文書処理装置を提供することを目的とする。

【０００８】

【課題を解決する為の手段】上述の課題を解決するために、本発明の文書フィルタリング方法は、受信した文書をフィルタリングして提示する文書フィルタリング方法において、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別し、該選別された文書を提示することを特徴とする。

【０００９】ここで、前記提示文書の選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する工程とを含む。また、前記閾値は更に文書数に応じて設定される。また、前記提示文書の選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記合致度と経過時間とに応じて受信文書にスコアを与える工程と、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する工程とを含む。また、前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされる。また、前記特徴ベクトルは、複数のキーワードまたは文から成る。また、前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方である。

【００１０】又、本発明の文書処理装置は、受信した文書をフィルタリングして提示する文書処理装置において、受信文書を作成からの経過時間及び／又は受信からの経過時間と共に保持する文書保持手段と、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別する選別手段と、該選別された文書を提示する提示手段とを備えることを特徴とする。

【００１１】ここで、前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記経過時間に対応する閾値を設定する閾値設定手段と、前記閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含む。また、前記閾値設定手段は、更に文書数にも応じて閾値を設定する。また、前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記合致度と経過時間とに応じて受信文書にスコアを与えるスコア計算手段と、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含む。また、前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされる。また、前記特徴ベクトルは、複数のキーワードまたは文から成る。

【００１２】

【作用】以上の構成により、文書フィルタリングの際に、文書特徴と選別特徴との合致度が、経過時間に応じて定まる閾値を越えた場合に、その文書を必要と判断してユーザに呈示することにより、文書の鮮度を考慮したフイルタリングを行うことが可能となる。

【００１３】

【実施例】以下、本発明の実施例を添付図面を用いて詳細に説明する。図１は、本実施例の文書処理装置の処理の論理構成を示すブロック図である。図１において、１０１は、ユーザの元に入ってきた文書を格納している文書データベースである。１０２は、処理対象の文書１０２－１と、その文書特徴１０２－２と、文書到着時刻１０２－３を保持する文書保持部である。１０３は、ユーザの関心に合う文書の文書特徴（選別特徴）を保持する選別特徴保持部である。１０４は、処理対象の文書の文書特徴と選別特徴との合致度を計算する合致度計算部である。１０５は、現在の時刻を保持する時刻保持部である。１０６は、文書が到着してからの経過時間を計算する経過時間計算部である。１０７は、経過時間をもとに閾値を計算する閾値計算部である。１０８は、合致度と閾値の関係によって文書を選択する文書選択部である。１０９は、文書選択部１０８で選ばれた文書を保持する選別文書保持部である。１１０は、選別文書保持部１０９に保持された文書を表示する文書表示部である。

【００１４】図２は本実施例の文書処理装置のハードウエア構成を示す図である。図２において、２０１は、図１及び図４に示す制御手順３００を記憶する制御メモリ

である。これはＲＯＭであってもよいし、ＲＡＭであっても良い。２０２は、制御メモリ２０１に記憶されている制御手段に従って処理を行う中央処理装置である。２０３はメモリで、上記文書保持部１０２，選別特徴保持部１０３，選別文書保持部１０９を有する。なお、時刻保持部１０５は現在の時刻を保持するところで、ハードウエアであっても良い。２０４はキーボード、２０８はポインティングデバイスであり、操作者が操作する。２０５はディスクであり、文書データベース１０１を有する。２０６はディスプレイで、ＣＲＴであってもよいし液晶ディスプレイであってもよい。これは文書を表示するのに用いる。２０７は各構成要素を接続する為のバスである。２０９はフロッピーである。

【００１５】図３は、図２に示す制御メモリ２０１の中にある制御手順３００の構成を更に示すものであり、合致度計算部１０４と、経過時間計算部１０６と、閾値計算部１０７と、文書選別部１０８と、文書表示部１１０とを含む。図４は図３に示した処理部に対応する動作手順を示すフローチャートである。図４を参照しながら、本発明の一実施例の動作を説明する。なお、本実施例では文書の特徴の表現方法として、一般に知られているベクトル空間モデルを利用する。

【００１６】ベクトル空間モデルでは、文書の特徴を表現するのに、Ｎ個のキーワードを用意し、文書毎に各キーワードの重みを設定する。これは、Ｎ次元空間のベクトルとみなすことができる。このベクトルを長さ１に正規化する。文書の特徴の合致度はそれぞれのベクトルの内積として表す。まず、ステップＳ３０１で、合致度計算部１０４において、文書保持部１０２に保持された文書特徴１０２－２と、選別特徴保持部１０３に保持された選別特徴との合致度を計算する。先に述べたように、合致度は文書特徴を表すベクトル間の内積で表すので、文書特徴をベクトルｄ，選別特徴をベクトルｓとすると、合致度は（ｄ・ｓ）になる。

【００１７】ステップＳ３０２では、経過時間計算部１０６において、文書が到着してからの経過時間を計算する。これは、時刻保持部１０５に保持された現在時刻と、文書保持部１０２に保持された文書到着時刻１０２－３との差分によって求めることができる。ステップＳ３０３では、閾値計算部１０７において、ステップＳ３０２で計算した経過時間から閾値を計算する。閾値は経過時間ｔに従って決めるが、ｔの増加と共に増加するような関数ｆ（ｔ）であればどのような決めかたをしても構わない。例えば、ｔを日数（端数切り捨て）で表現し、ｆ（ｔ）＝１－１／（ｔ＋２）のようにする。

【００１８】ステップＳ３０４では、ステップＳ３０１で計算した合致度と、ステップＳ３０３で計算した閾値を比較する。合致度が閾値を越えていない場合はそのまま処理を終了する。ステップＳ３０４で合致度が閾値を越えていればステップＳ３０５に進み、文書選別部１０８で、この文書を選択して選別文書保持部１０９に保持する。そして処理を終了する。

【００１９】例えば、ある文書の文書特徴ベクトルと選別特徴ベクトルの内積ｐを０．７８とする。この文書が到着してから２日経過した時点ではｆ（ｔ）＝１－１／（ｔ＋２）に代入すると、ｆ（２）＝０．７５であり、ｐ＞ｆ（２）となって、文書を選択するが、３日経過した時点ではｆ（３）＝０．８であり、ｐ＜ｆ（３）となって、文書を選択しないことになる。

【００２０】次に、文書表示部１１０により、選択された文書を表示する。尚、前記実施例では、文書をユーザに呈示する際に合致度と閾値の計算を両方行っているが、合致度の計算を文書が届いた時点で行ない、その値を保存しておくようにしておき、次に、文書をユーザに呈示する際には閾値の計算だけを行うようにする。こうすることによって、文書をユーザに呈示する際の処理時間を短縮することができる。

【００２１】又、前記実施例では、文書の特徴にベクトル空間モデルを利用したが、この表現方法に限らない。すなわち、文書特徴と選別特徴の合致度を計算し、それに対する閾値を経過時間ｔの関数として定義できれば良い。又、前記実施例では閾値を計算する際に経過時間を日数で表わしたが、秒・分・時間などどんな単位で表現しても構わない。また閾値の関数も上記実施例に挙げたものに限るものではない。

【００２２】又、前記実施例では、経過時間ｔの関数で定まる閾値を用意して合致度との大小を調べたが、文書の選別特徴の合致度と経過時間ｔから定まるスコアの合計を文書選別のスコアとし、文書選別のスコア自体に文書特徴の合致度だけでなく、経過時間ｔも反映するようにし、更に、経過時間ｔの増加でスコアが減少するように定めれば良い。このスコアに対して一定の閾値を設けて文書を選別するものでもよい。

【００２３】又、前記実施例では、溜まった文書数を考慮していない。しかし、閾値を計算する際に文書数も反映して、文書数が多い場合にはそれだけ閾値を高くするようにしても良い。こうすることによって、文書数が多い場合でも一定の数の文書をユーザに提示することができるようになる。又、前記実施例では、文書が到着してからの経過時間を用いたが、文書が作成された時刻がわかる場合は、作成されてからの経過時間を用いるようにしても良いし、作成と受信とを共に用いてもよい。

【００２４】更に、本発明は、複数の機器から構成されるシステムに適用しても、１つの機器から成る装置に適用しても良い。また、本発明はシステム或は装置にプログラムを供給することによって達成される場合にも適用できることはいうまでもない。

【００２５】

【発明の効果】以上説明したように、本発明によれば、より新しい情報をより重要なものとして文書フィルタリ

ングを行うようにしたので、限られた時間でユーザが情報を効率的に利用することができるという効果がある。

【図面の簡単な説明】

【図１】本実施例の文書処理装置の処理の論理構成を示すブロック図である。

【図２】本実施例の文書処理装置のハードウエア構成を示す図である。

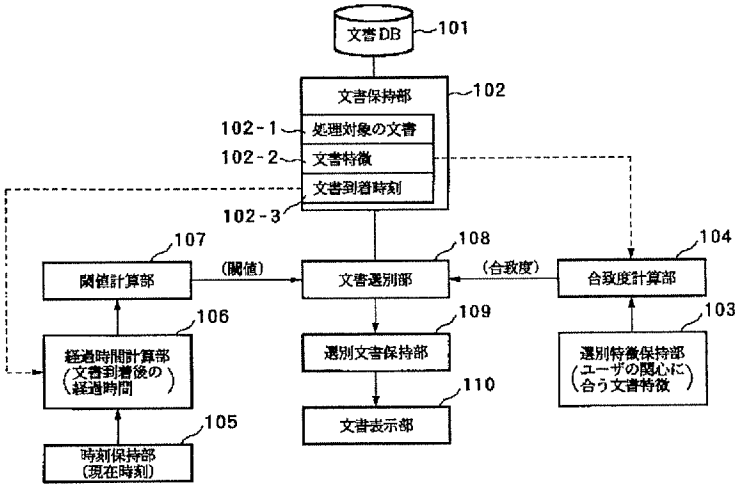
【図３】本実施例の制御手順の構成を示す図である。

【図４】本実施例に係る動作手順を示すフローチャートである。

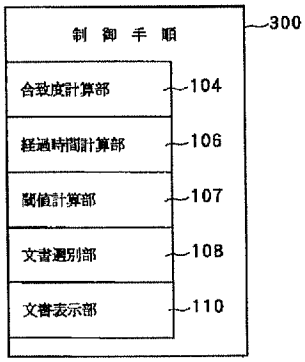
【符号の説明】

１０１ 文書データベース
１０２ 文書保持部
１０２－１ 処理対象の文書
１０２－２ 文書の特徴
１０２－３ 文書到着時刻
１０３ 選別特徴保持部
＊１０４ 合致度計算部
１０５ 時刻保持部
１０６ 経過時間計算部
１０７ 閾値計算部
１０８ 文書選別部
１０９ 選別文書保持部
１１０ 文書表示部
２０１ 制御メモリ
２０２ 中央処理装置
２０３ メモリ
２０４ キーボード
２０５ ディスク
２０６ ディスプレイ
２０７ バス
２０８ ポインティングデバイス
２０９ フロッピー

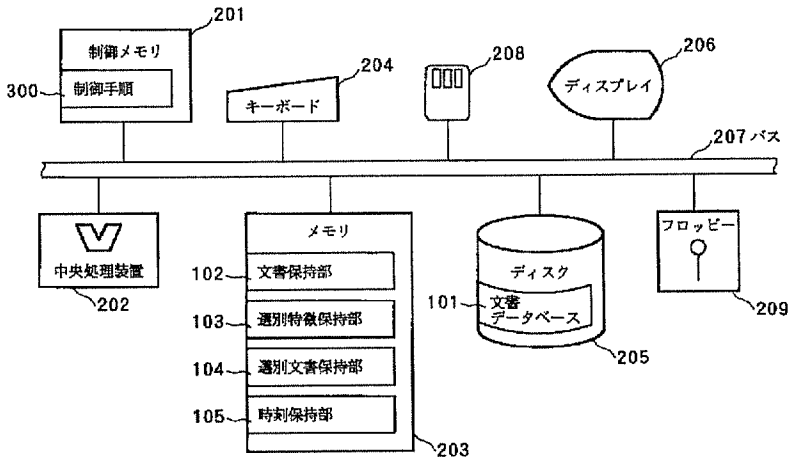
【図１】



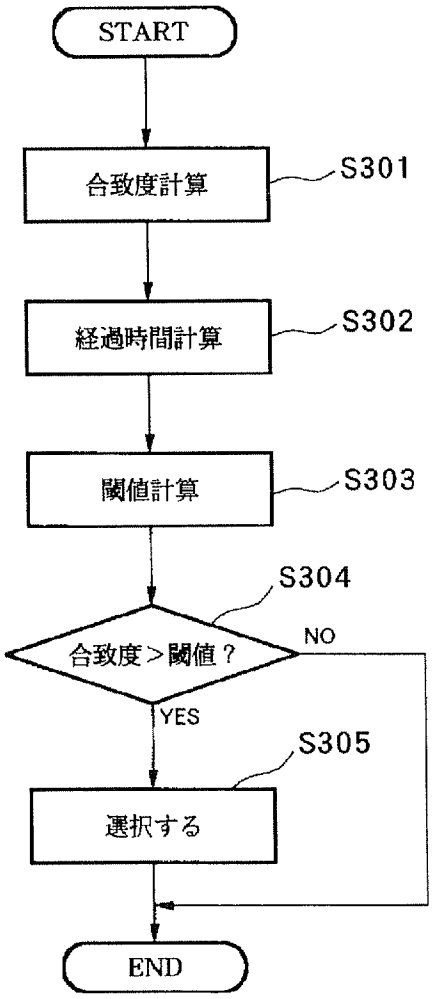
【図３】



【図２】


201
制御メモリ
300
制御手順
204
キーボード
208
206
ディスプレイ
207 バス
中央処理装置
202
メモリ
102
文書保持部
103
選別特徴保持部
104
選別文書保持部
105
時刻保持部
203
ディスク
101
文書
データベース
205
フロッピー
209

【図４】



フロントページの続き

(51)Int.Cl.[6]	識別記号	庁内整理番号	F I		技術表示箇所
		9194－5L		15/40　３１０　F	

(72)発明者　柴田　昇吾
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　池田　裕治
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　藤田　稔
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内